




## ARTICLE

# Automating the Proposition of Neologisms for the Quechua Language

Luis Camacho 

Pontificia Universidad Católica del Perú  
Email: [camacho.l@pucp.pe](mailto:camacho.l@pucp.pe)

(Received 31 July 2023; revised 2 October 2024; accepted 4 October 2024; first published online 2 May 2025)

## Summary

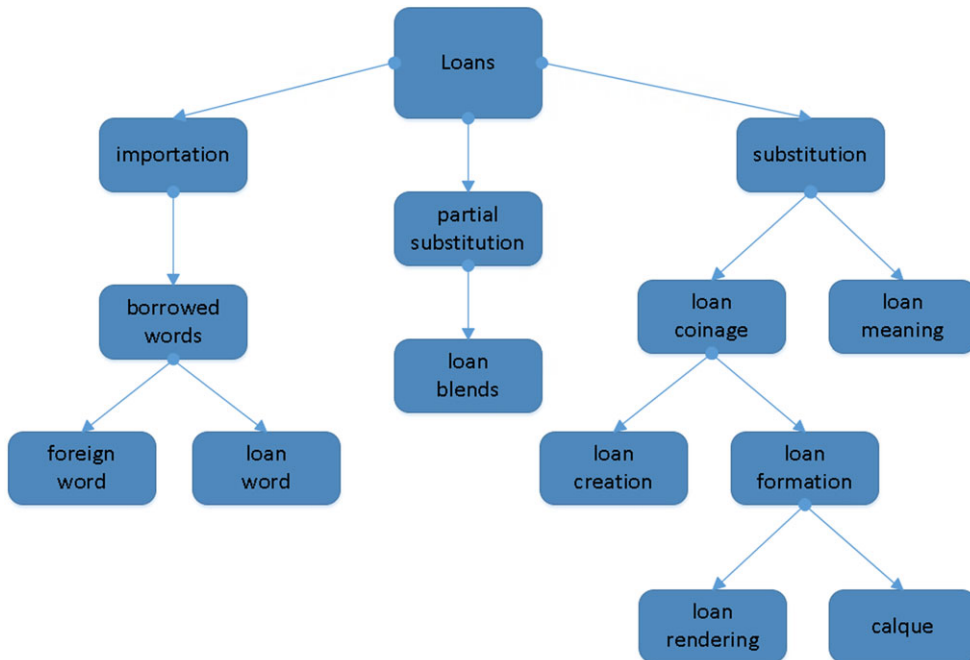
I introduce a discussion about the importance of coining new words for endangered languages. Coining new words is one of the many tasks that must be done to revitalize endangered languages. I propose a method for automating the process of coining new words and evaluate the method in the Quechua language. The method starts by collecting a list of English words, then translating these words into several other languages, generating a list for each language, and performing IPA notation of all lists; finally, a rule-based algorithm identifies words that match the phonotactics of the target language. The method can propose thousands of words as neologisms in the target language.

**Keywords:** lexical creativity; quechua; under-resourced languages; phonotactics; loanwords

## 1. Introduction

Language Policy and Planning (LPP) is a specialized domain within sociolinguistics, focusing on the intricate interplay of societal norms, beliefs, and actions that mold language usage within communities, alongside deliberate efforts to mold and regulate language practices (García 2015). Nahir (Nahir 1984) delineates 11 objectives within language planning, with lexical modernization ranking among them. Ahlers & Hinton (Ahlers & Hinton 1999) identify borrowing and word formation as the principal mechanisms for novel lexical items entering a language. Sager (Sager 1990) distinguishes between two primary categories of new lexical entities: those that are entirely novel creations and those that are borrowed from other languages. The former category, called coinages, represents a unique form of neologism, wherein a previously nonexistent sound sequence is imbued with meaning, thus forging a fresh sound-meaning pairing (Akmajian et al. 2010). Haspelmath (Haspelmath 2009) defines lexical borrowing, or loanword, as a word that, at some point in the history of a language, entered its lexicon as a result of borrowing (or transfer or copying). Further differentiation is made by Akmajian between direct borrowings, entailing the adoption of sound and meaning, and indirect borrowings, encompassing loan translations or calques, which involve the literal translation of words from another language.

Loanwords undergo diverse adaptation processes when transitioning from one language to another. Werner Betz (Betz 1949), Einar Haugen (Haugen 1950), and Uriel Weinreich



**Figure 1.** Classification of borrowings.

(Weinrich 1953) made seminal contributions to the theoretical understanding of loan-word influence (Oksaar 1996). The foundational theoretical frameworks largely derive from Betz’s nomenclature, although Duckworth enhanced Betz’s system by introducing the concept of “partial substitution” (Grzega 2003). Haugen further refines the categorization of borrowings into three fundamental groups based on the importation-substitution dichotomy:

1. importation, refers to the process of adopting words from another language into the target language without making any changes (or few) to their form or meaning
2. partial substitution, involves replacing some elements of the borrowed word with native elements while retaining other parts of the original word, this can happen at the phonetic level (sounds), morphological level (word forms), or semantic level (meanings)
3. substitution, is the process where the original phonological, morphological, or semantic elements of a loanword are completely replaced with native elements from the borrowing language. The concept or object the loanword refers to is maintained, but expressed entirely through the linguistic resources of the borrowing language.

A schematic illustration of this classification is shown in Figure 1.

The language planning challenges faced by endangered languages diverge significantly from those of mainstream languages. It encompasses more than merely creating new words; it delves into complex issues such as decolonization, reclaiming autonomy, and affirming indigenous identity. The imposition of dominant languages, such as English, French, or Spanish, on indigenous languages has led to the forced abandonment of these languages, resulting in a shift towards the dominant language. For sure, lingua francas play

a role in global communication, but we point out the pernicious effect of the diglossia that has prevented indigenous people from creating more lexicon in their languages, a serious issue since the start of the industrial revolutions. These multifaceted problems are not only intricate but also tend to provoke diverse perspectives and opinions.

Unlike widely used languages, endangered languages frequently lack terminology in critical domains such as education, various sciences, and politics. This shortfall arises because these languages have not been utilized for communication within these areas, leading to an undeveloped vocabulary. Consequently, the requirement for new terminology can be substantial. Over time, languages naturally evolve, yet some individuals believe in the language's perfection in its current state, viewing any deliberate changes as undesirable. This notion of preserving a single, authentic form of language, free from external influences, is known as linguistic purism (Sallabank & King 2021). The attitudes and ideologies of those advocating for language revitalization, as well as the wider community, significantly influence the success or failure of these efforts, including the modernization of the lexicon (Coronel-Molina 2015). A purist stance towards language revitalization often manifests in resistance to loanwords, particularly from dominant languages, which is linked to concerns over language loss (Blair & Fredeen 1995). For instance, the Sámi in Scandinavia have undertaken efforts to preserve and revitalize their language, facing challenges of language and culture loss due to long histories of assimilation policies. The development of new vocabulary to cover modern concepts while keeping the language's integrity reflect a resistance to Norwegian, Swedish, or Finnish borrowings (Aikio-Puoskari 2018).

Coining of new words is a natural part of language evolution. As the world changes, so too do languages. The new words coined today will help to shape the language of tomorrow, and that is true for all languages: dominants, endangered, under-resourced, or indigenous ones. There is no definitive way to count the number of new words coined each year, as no single source tracks all new words (Algeo & Algeo 1991). In addition, the definition of what constitutes a "new word" can vary. Some people might consider a new word to be any word not found in a dictionary, while others might only consider it new if it is widely used. According to the *Oxford English Dictionary* (OED), the number of new words they added to their dictionary each year has varied since 2000. In the early 2000s, the OED added 1,000 new words annually on average. However, the number of new words added to the dictionary has increased in recent years, with an average of 1,500 new words added annually between 2010 and 2020, totalling up to 25,000 new words this century. The evidence suggests that, as part of revitalization efforts, speakers of endangered languages must actively engage in **lexical creativity**. This involves not only keeping pace with the evolving linguistic landscape but also **coining new words** that reflect their unique lived experiences. Furthermore, it is essential for these speakers to popularize these terms globally, much like how neologisms from dominant languages achieve mainstream recognition. In that direction, this paper builds upon and enhances previous work (Camacho 2023) on automating the generation of neologisms for the Quechua language. The contributions are:

1. an improved script to select candidate neologisms
2. a longer list of candidate words from the previously used source, Open Dictionary<sup>1</sup>, now there are 33,818, all translated to English and Spanish except 2,384 Cantonese words
3. an expanded list of candidate words from the previously used source, Wikipron (Jackson et al. 2020); there are now 23,236, with 14,577 translated into both English and Spanish

<sup>1</sup> <https://github.com/open-dict-data/ipa-dict>

4. the introduction and analysis of a third data source: a list of 370,107 English words<sup>2</sup> translated into 57 languages and from all these languages, 186,322 proposed neologisms
5. a detailed description of dealing with processing text using modern tools like Google Cloud API Translate and Python libraries

## 2. History of lexical modernization of the Quechua language

Haimovich & Szemiński 2018 conducted a historical survey of Quechua language planning in Peru and Bolivia, summarized here. From the nineteenth century to the late twentieth century, various factors influenced Quechua's development. Early efforts were marked by inconsistency. After a period of suppression, Quechua was employed in the Independence wartime (1815–1825) propaganda, often using archaic terms and newly created words that were difficult for everyday speakers to understand. Peruvian authorities showed little interest in systematic language planning. After the devastating war against Chile (1879–1884), two contrasting ideologies emerged: the “educationalist view” which advocated for the assimilation of Quechua speakers into Spanish-speaking society through education, and “liberal indigenism” which emphasized the preservation and revitalization of the Quechua language.

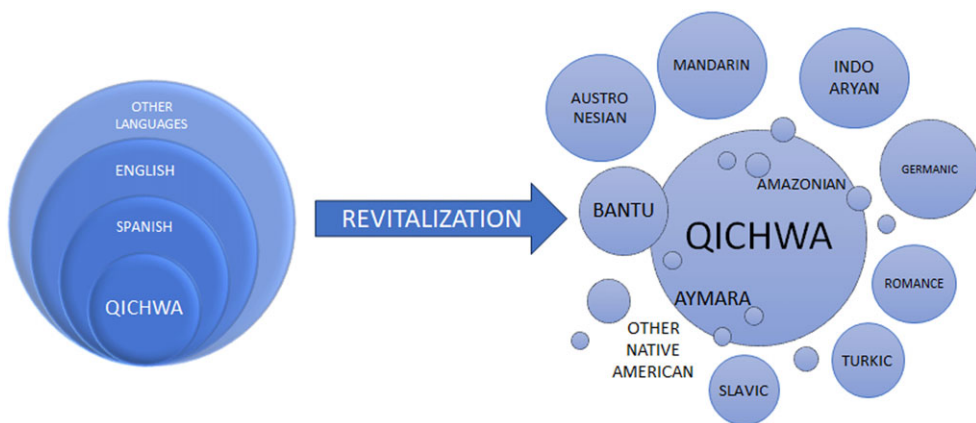
The early twentieth century saw Cusco become a center for Quechua language planning. Here, liberal indigenists sought to create a “great” variety of Quechua so-called “Qhapaq Simi.” They aimed to purify the language by removing Spanish influences and incorporating words from colonial sources. However, this approach resulted in a form of Quechua quite different from the everyday language spoken by the common people. Mid-century developments included the Academy of the Quechua Language (ALQ) foundation in Cusco, promoting Qhapaq Simi and language purity with limited impact on the general public (Coronel-Molina 2007). During this period, bilingual education programs were launched in Quechua-speaking regions of Peru. However, these programs focused on transitioning students to Spanish rather than fully developing Quechua literacy.

A significant milestone came in 1975 when Quechua became an official language of Peru alongside Spanish. However, the specific variety of Quechua to be used and the implementation plans remained unclear. Worst of all, the eurocentric and monolingual Spanish-speaking elite opposed the law. Finally, the act was downgraded from national to regional level and, even worse, never enforced. Bilingual education programs (IBE) were launched in the late twentieth century, with a focus on interculturality. Professional linguists also played an increasingly important role in IBE projects and Quechua's development to meet its diverse speakers' needs. Several proposals were put forth to establish a unified standard for the written language, with the one proposed by Cerrón Palomino ultimately being selected (Cerrón-Palomino 1990). In 1985, the Ministry of Education of Peru issued the order 1218-85-ED, setting the official alphabet of Quechua (and also of Aymara), which is still in force.

## 3. Previous work

Considering the background above, I have focused on solving four of the problems present in the lexical modernization of the native languages from South America:

<sup>2</sup> <https://github.com/dwyl/english-words/>



**Figure 2.** De facto and desirable status of the Quechua language in contact with other ones.

1. the growing gap concerning the dominant languages' lexicon
2. the constraints on the use of own resources
3. the need to adapt loanwords to the rules of the receiving language
4. Spanish as a source of loans

The increase in the number of new words added to the dominant languages' vocabularies in recent years is likely due to many factors, including the rapid pace of technological change, the increasing globalization of the world, and the growing popularity of social media. New technologies often give rise to new terms, as people come up with new ways to describe the things that they see and do.

In both Peru and Bolivia, a discourse surrounds the utilization of loanwords from external languages in the modernization efforts of Quechua. Cerrón Palomino advocates for the preference of neologisms crafted from indigenous sources as a more favorable approach to enhancing Quechua. Moreover, many minority and indigenous languages commonly rely on descriptive constructs for word formation. These descriptions often delineate physical attributes, actions, or purposes, varying in their classification as compounds or phrases across languages. While some of these extended expressions may evolve into recognized words, an overreliance on this method can pose challenges. Therefore, while the length of newly coined terms is not inherently problematic, excessive use of descriptive circumlocutions should be avoided (van der Voort 2016).

Many times loanwords are not accepted for more social and political than technical reasons. These difficulties result from the diglossic situation in which endangered languages are found; in other words, European colonialism imposed European languages at the top of a hierarchy to the detriment of the native languages and the wealth of knowledge of their speakers. The end of colonialism did not change this dynamic; as a result, many speakers of indigenous languages find themselves surrounded by the language of the colonizers as shown in Figure 2, which impedes them from establishing direct bilateral relations with other languages and cultures but forces them to use the dominant language as a compulsory bridge.

In the case of South America, objections are not to the foreign source of new words but to the imposition of neologisms from the Spanish language, a legacy of colonialism. However, in the same way that when a behavior is normalized by a dominant cultural environment, it becomes invisible, most of the Peruvian State seems not to realize this, which is noticeable

in the bilingual education programs and publications of the Peruvian Ministry of Education (Huamancayo 2017). Last but not least, the challenge is that if terms are borrowed from other languages, they should at least be accommodated to Quechua phonotactics.

While there is a preference for using native sources for neologisms, loanwords can be used if they are adjusted to fit the Quechua language, are proposed and not imposed, and do not reinforce diglossia. In that direction, I established a new paradigm: borrowing words which meet the Quechua phonotactics regardless of the language of origin of such words. In my previous article (Camacho 2023), I introduced a novel approach, a computerized search of proposed loanwords for the Quechua language. The foundation of my research is phonotactics. In essence, phonotactics, a branch of phonology, is the study of rules and restrictions related to the arrangement of sounds within the syllables of a language, crucial for fluent and smooth pronunciation. It encompasses assessing how phonemes can be combined in a language, with phonotactic constraints specific to each language (Vitevitch & Luce 1999). Since it is difficult and expensive to analyze sound files, I relied on parallel corpora with IPA transcriptions. I split the tasks into three parts: finding sources, searching inside the sources for eligible words, and finally translating those words. I found two sources of IPA representation of several languages: Open Dictionary and Wikipron. I wrote a Python script that incorporated all the spelling and pronunciation rules of the Quechua language. After executing the script, two intermediate files collected all the words with the requested Quechua phonotactics. The two intermediate files were inputs to another script that identified the words and their respective languages present in the intermediate files; if the script finds both in WordNet (Fellbaum 2017), it delivers the English translation. To be exact, the English synset was searched for in WordNet, and all the lemmas associated with that synset were listed. This way, from the Open Dictionary, we found 14,970 (0.39% of all words examined) neologisms and 1,768 (11.81%) were translated into English. From Wikipron, 26,752 (0.76% of the total amount) neologisms were found, and then 6,118 (22.87%) were translated into English.

The low percentages of eligible words selected by the algorithm suggested that it was necessary to adjust the scripts, overall improving data cleaning and coding diacritics. I examined seven million words, but this is just a small fraction of the universe of words in all languages; then the challenge is to find more sources and to make the outputs useful. Authoritative translation sources and translation of words to the English language are also needed. I deal with these three tasks in this present article.

#### 4. Adding sources

I analyzed several sources looking for parallel corpora of words and IPA transcriptions in as many languages as possible, like the UCLA phonetic corpus (Li et al. 2021), the Vox Communis Corpus (Ahn & Chodroff 2022), and the CMU Wilderness Multilingual Speech Dataset (Black 2019).

The UCLA phonetic corpus is a collection of audio recordings and wordlists of speech from 97 languages. It contains recordings of male and female speakers and includes a wide range of dialects and accents. The recordings are made at a high sampling rate and are accompanied by detailed phonetic transcriptions. There are around 70 words per language; however, there is no translation to English or scripting of the words in their writing systems.

The Vox Communis corpus contains acoustic models, lexicons, and force-aligned TextGrids with phone and word-level segmentations derived from the Mozilla Common Voice corpus. The Mozilla Common Voice corpus contains audio data with transcriptions from over 30 languages. The lexicons are developed using Epitran (Mortensen et al. 2018)



and the XPF Corpus, both rule-based G2P systems. The acoustic models have been trained using the Montreal Forced Aligner (version 2.0), and the force-aligned TextGrids are obtained directly from those alignments. These acoustic models can be downloaded and re-used with the Montreal Forced Aligner for new data. The lexicons are aligned with IPA notation, but the corpus is small.

The CMU Wilderness Multilingual Speech Dataset is a dataset of over 700 languages providing audio, aligned text, and word pronunciations. On average, each language provides around 20 hours of sentence-length transcriptions. However, the corpus does not contain IPA notation, and the data is mined from reading New Testaments<sup>3</sup>.

Overall, I found that none of these sources met all our requirements. The UCLA phonetic corpus is the most comprehensive but does not include IPA notation or scripting of the words in their writing systems. The Vox Communis corpus does include IPA notation, but it is relatively short. The CMU Wilderness Multilingual Speech Dataset is the largest, but it does not include IPA notation, and the data doesn't come from various sources.

As a result, I opted to modify the approach. Instead of seeking parallel corpora for text and IPA notation, I prioritized obtaining a comprehensive list of English words from a reputable and freely available source. I then utilized the Google Cloud Translate API to translate these words into languages with an automatic IPA converter available. Initially, I focused on WordNet and BabelNet as primary sources. WordNet is more English-centric, with a narrower scope, and primarily expert-driven. At the same time, BabelNet covers a wider range of languages, offers extensive coverage, and combines expert knowledge with large-scale automatic extraction from web sources. However, BabelNet's programmatic approach is complex, and WordNet lacks sufficient data, with a limited number of languages (32) and a shortage of words per language.

Finally, I found the solution with Epitran, a library and tool for transliterating orthographic text as IPA for 115 languages; 57 of these languages can be translated with Google Translate. I also found English Words<sup>4</sup>, a source of 370,107 words. An advantage of this approach is knowing the meanings of all proposed neologisms, which was not possible in the previous article, where the sources were Open Dictionary and Wikipron.

## 5. Improving the selection algorithm

In the previous article, I found all the proposed neologisms were morphologically correct, demonstrating the accuracy of the selection algorithm's encoding rules. Surprisingly, only 0.57% of the over seven million words examined were selected, likely due to the precision level of the sources. The sources offered broad and narrow transcriptions, with narrow transcriptions containing more detailed phonetic information. While narrow transcription aids in producing precise sounds and allows for more detailed analysis, it may not represent all speakers. It could be challenging for non-specialists due to its larger number of unfamiliar symbols and diacritics. Broad transcription, on the other hand, is suitable for making statements applicable across diverse language communities and is commonly used in foreign language dictionaries. The sources used in the study contained both types of notations, but diacritics were not encoded, resulting in the limited selection of eligible words. Even worse, many IPA symbols, not only those with diacritics, are represented by double characters, which was not noticed in the previous article. I also found confusion in representing ejective consonants where one of the sources used (') instead of the right symbol ⟨⟩; we noticed that confusion even happens with the lateral release notation which is the release

<sup>3</sup> <http://www.bible.is/>

<sup>4</sup> <https://github.com/dwyl/english-words/>

of a plosive consonant into a lateral consonant and it is transcribed in the IPA notation with a superscript (l).

That problem was solved with more secure tokenization using the Python package IPAtok (Li et al. 2021). IPAtok takes an IPA string and returns a list of tokens. A token usually consists of a single letter together with its accompanying diacritics, but if two letters are connected by a tie bar, they are also considered a single token. Except for length markers, suprasegmentals are excluded from the output, and whitespaces are ignored. I decided to use IPAtok not only with the first sources but, going forwards, for any source of words.

The lack of cleanliness in the data caused the second problema, so it was necessary to clean manually but, above all, to improve the code so that it could bypass any uneven piece of data, instead of trying to fix it, without stopping the process of selecting candidate words.

## 6. Pipeline

Under the new paradigm, the process follows the next steps:

1. it started from a large but simple file, TXT or CSV, containing words in some language. In this article, we started with English
2. with this list of English words we used the Google Cloud Translate API to create lists of those words translated to all the languages that we can deal with in the next step
3. we used Epitran to translate to IPA this whole list of words; right now Google Translate and Epitran support only 57 languages in common
4. tokenization of all IPA notations using IPAtok (Sofroniev & Çöltekin 2018)
5. selection of all the eligible neologisms using our algorithm
6. report of findings

As described previously, I selected English Words, a list of over 466,000 single words pulled out into a simple new-line-delimited text file. Actually, I selected a shorter version, a list of 370,107 words that excludes symbols, numbers, proper names, acronyms, or compound words and phrases but does not exclude archaic words or significant variant spellings.

The second step was translation. Creating a Python script speeds up the machine translation process from the start. Multiple Python packages offer these capabilities, but except for Google's official API, all other packages are neither stable nor supported by Google. Hence, the more efficient way to use Google Translate for machine translation in Python projects is Google's official API. Google provides two different versions of the Cloud Translation API<sup>5</sup>; to keep things simple we selected the basic version. Using Google Translation API there are four steps: cloud setup<sup>6</sup>, API setup, installation, and implementation. First, you must create a project via Google Cloud Console. You will need a project number or ID when calling the Translation API. Then, in your local environment, you must enable the Google Translate API by running the following command:

```
gcloud services enable translate.googleapis.com
```

then, it's recommended to create a Python virtual environment to isolate the dependencies:

```
virtualenv venv-translate
```

<sup>5</sup> <https://cloud.google.com/translate#all-features>

<sup>6</sup> <https://cloud.google.com/translate/docs/setup>



activate the virtual environment:

```
source venv-translate/bin/activate
```

next, we installed the Translation API client library:

```
pip install ipython google-cloud-translate
```

finally, create a new Python file and add the following import statement at the top of the file:

```
from google.cloud import translate_v2 as translate
```

next, create a new function, and inside it, instantiate a `translate.Client()` object as follows:

```
translate_client = translate.Client()
```

At first, I tried to run the script on a local laptop but it stopped working once and once again, and even worse, the estimated translation time per language was more the 24 hours; so I decided to use Google Cloud instead. I split the 370,107 words list into three segments and ran translations of six files at the same time (that means two languages at the same time). This way, to translate to 58 languages, I had to process 174 files, three files per language, and that took seven days, five hours, and 36 minutes and cost around 500.

The third step was key, transcribing the words in the 59 languages (including English) into IPA notation. For this, I used Epitran, a multilingual, multiple back-end system for G2P (grapheme-to-phoneme) transduction. It takes word tokens in the orthography of a language and outputs a phonemic representation in either IPA or X-SAMPA. I installed Epitran by executing:

```
pip install Epitran
```

For use with most languages, Epitran requires no special installation steps. However, English G2P in Epitran relies on CMU Flite, a speech synthesis package by Alan Black and other speech researchers at Carnegie Mellon University. For the current version of Epitran, the recommended way is to obtain the Flite source from GitHub<sup>7</sup>, uncompressed, and compile the source. To use the Epitran class for Mandarin Chinese (Simplified and Traditional) G2P, it also needs additional steps; it is necessary to point the constructor to a copy of the CC-CEDict<sup>8</sup> dictionary, then run one time:

```
import Epitran
epi = Epitran.Epitran('cmn-Hans', cedit_file='cedict_1_0_ts_utf-8_mdbg.txt')
```

I ran Epitran on a local laptop, and it took around 40 minutes to create the IPA notation of the 59 languages' words. To avoid false negatives in the selection of proposed neologisms, the next step was the tokenization of the IPA notations for what I installed IPAtok:

```
pip install ipatok
```

and it is called in a script with

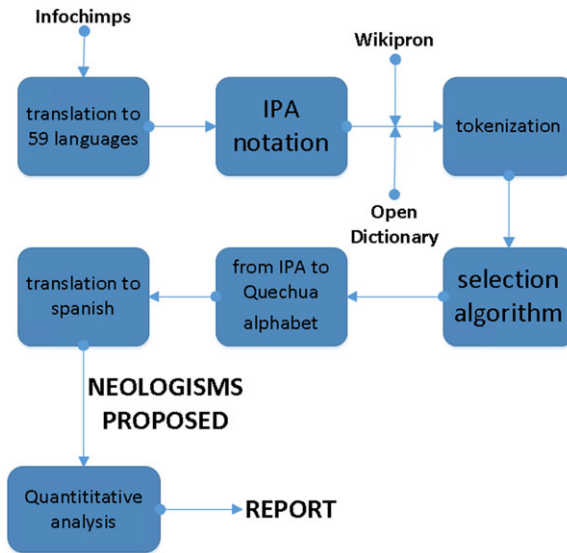
```
from ipatok import tokenise
```

Then I put all these inputs in an improved neologisms selection script, and the output was the list of eligible words. Finally, the selected IPA notations were translated into the Quechua script, and a results report was elaborated on. This last process took more than two hours to run. The code is available in GitHub<sup>9</sup>. The full process is illustrated in Figure 3.

<sup>7</sup> <https://github.com/festvox/flite>

<sup>8</sup> <https://cc-cedict.org/wiki/>

<sup>9</sup> <https://github.com/luisacamachocaballero/QuechuaNeologismsFromDemand/blob/main/SelectingNeologism4QuechuaUsingGoogleCloudAPItranslateIPAtokEpitran.ipynb>



**Figure 3.** Architecture of the system.

## 7. Results

Results are shown in Table 1 and Table 2. From the new source, English Words, 17,691,783 words, corresponding to 59 languages, were examined, of which 186,322 (1.05%) neologisms were found. All the languages contributed, from Swahili, which contributed 17,305 (4.67% of Swahili examined words) proposed neologisms, to Arabic, which contributed with just four (0.007%). Twenty-one languages contributed with more than 1% of the words examined in such languages. The improvement of the selection code also led to an increase in the number of proposed neologisms from the two previous sources.

The first source, Open Dictionary, has 3,838,348 words, of which 14,970 (0.39%) neologisms had been found, from which 1,768 (11.81%) were translated into English; now, from the same source, I found 33,818 (0.88%) candidate neologisms, all translated (92.95%) to English and Spanish except 2,384 Cantonese words. In absolute terms, more contributions came from the Spanish language (4,135 words), but in relative terms, the contribution of the Swahili language is impressive (7.09%; 3,425 words). The contributions of the languages Mandarin, Cantonese, and Vietnamese are also remarkable, in absolute and relative terms, since previously these languages had null contributions.

The second source, Wikipron, has 3,506,258 words, of which 26,752 (0.76%) had been found, and then 6,118 (22.87%) were translated into English; now there are 23,236 (0.66%) proposed neologisms of which 14,577 were translated (62.73%) to English and Spanish. In absolute terms, the top ten contributors are Mandarin Chinese (2,857), Min Nan Chinese (1,773), ancient Greek (1,170), Polish (988), Spanish (980), Vietnamese (928), French (914), Latin (826), Catalan (799), and Finnish (568), but none of these languages overcome the 4% in relative terms. In the relative realm, the story is different, with 20 less-spoken languages having the higher figures including Carrier, a North American language, as the major relative contributor with 23.43% (41 words) and also Hawaiian with 15.18% (95 words). A priori, I should conduct deeper research in those languages in which I have achieved high percentages; however, it should also be taken into account that there are two phenomena. The first is that the highest percentages are found in languages that are in a similar situation to the Quechua language and therefore are probably also searching for neologisms for more

**Table 1.** Results from the Open Dictionary source

		Eligible Words		Eligible/Total	
Language	Total Words	before	now	before	now
Swahili	48,308	2,544	3,425	5.27%	7.09%
Mandarin traditional	48,240		3,363	0.00%	6.97%
Mandarin simplified	44,780		3,017	0.00%	6.74%
Jamaican Creole	1,870	73	79	3.90%	4.22%
Cantonese	56,826		2,384	0.00%	4.20%
Malay (Malaysian and Indonesian)	28,215	693	763	2.46%	2.70%
Vietnamese (Southern)	72,234		1,790	0.00%	2.48%
Vietnamese (Central)	72,234		1,430	0.00%	1.98%
Vietnamese (Northern)	72,234		955	0.00%	1.32%
French (Québec)	245,958	2,526	2,533	1.03%	1.03%
English (General American)	125,927	3	1,245	0.00%	0.99%
French (France)	245,178	1,969	1,989	0.80%	0.81%
Persian	8,090	60	61	0.74%	0.75%
Spanish (Mexico)	595,885	2,903	4,135	0.49%	0.69%
Spanish (Spain)	595,896	2,903	4,135	0.49%	0.69%
Finnish	92,837	21	643	0.02%	0.69%
Esperanto	23,517	155	157	0.66%	0.67%
Norwegian Bokmål	10,172	38	60	0.37%	0.59%
Japanese	221,421	1,005	1,018	0.45%	0.46%
English (Received Pronunciation)	65,118		292	0.00%	0.45%
Odia	6,226	23	26	0.37%	0.42%
Arabic (Modern Standard)	857,161	23	250	0.00%	0.03%
German	278,915	28	66	0.01%	0.02%
Swedish	21,106	3	2	0.01%	0.01%
	<b>3,838,348</b>	<b>14,970</b>	<b>33,818</b>	<b>0.39%</b>	<b>0.88%</b>

modern concepts. Secondly, these same languages have had a small list of examined words, so extrapolating that many more neologisms will be found with larger lists is venturing too far. It is noticeable that the Swahili language was not included in this source. The three lists of proposed neologisms, one per source, are available in our repository<sup>10</sup>.

<sup>10</sup> <https://github.com/luiscamachocaballero/QuechuaNeologismsFromDemand/tree/main/outputs>

**Table 2.** Comparison of results from Wikipron and English Words sources

Language Identification		Wikipron Source				English Words Source			
		Examined Words		Eligible Words		Eligible/Examined			
ISO 639-3	Language Name	before	now	before	now	before	now	elegibles/ examined	elegibles examined
swa	Swahili							4.67%	17,305 370,708
mri	Maori							4.24%	15,725 370,887
tgl	Tagalog	16,359	8535	544	482	3.33%	5.65%	3.20%	11,506 359,100
zul	Zulu	1,733	1,733	1	3	0.06%	0.17%	2.70%	9,998 370,650
nya	Nyanja	825	825	6	32	0.73%	3.88%	2.15%	7,956 370,479
cmn	Chinese	133,686	133,686	2	2,857	0.00%	2.14%	2.13%	7,667 359,679
jav	Javanese							1.60%	5,654 353,174
hau	Hausa	3,381	1,703	0	49	0.00%	2.88%	3.54%	5,558 156,965
hmn	Hmong							1.54%	4,868 315,920
cat	Catalan	66,099	66,099	765	799	1.16%	1.21%	1.30%	4,815 371,470
tur	Turkish	7,287	3,591	179	100	2.46%	2.78%	1.30%	4,806 370,702
msa	Malay	4,665	3,015	263	214	5.64%	7.10%	1.49%	4,657 311,572
spa	Spanish	347,921	58,883	5,098	980	1.47%	1.66%	1.22%	4,554 373,648
kin	Kinyarwanda							2.14%	4,459 208,294
yor	Yoruba	2,636	2,636	0	157	0.00%	5.96%	1.14%	4,212 370,799
uig	Uighur	530	270	40	20	7.55%	7.41%	1.36%	4,029 296,285
sqi	Albanian	2,318	1,484	81	48	3.49%	3.23%	1.12%	3,945 351,635
xho	Xhosa	871		0		0.00%		1.92%	3,937 204,997
mya	Burmese	9,811	4,909	0	105	0.00%	2.14%	0.91%	3,380 370,110
sin	Sinhala							1.51%	3,348 221,598
tam	Tamil	7,538	3,786	33	28	0.44%	0.74%	0.94%	3,174 339,027
pan	Panjabi	452	452	4	4	0.88%	0.88%	0.98%	2,918 297,298
fra	French	124,377	62,155	1,720	914	1.38%	1.47%	0.75%	2,827 376,859
	Other 36 languages	1,445,083	889,321	7,190	4,869	0.50%	0.55%	0.44%	45,024 10,199,927
	Other 197 languages	1,330,686	770,597	10,826	11,575	0.81%	1.50%		
<b>TOTAL</b>		<b>3,506,258</b>	<b>2,013,680</b>	<b>26,752</b>	<b>23,236</b>	<b>0.76%</b>	<b>1.15%</b>	<b>1.05%</b>	<b>186,322 17,691,783</b>

## 8. Discussion

In principle, the main finding is the large number of foreign words that do not need rephonologization to be inserted into the Quechua language as neologisms; the new source presented in this research added 186,322 candidate words. When added to the candidates from the other two sources, the number rises to 223,376 proposed neologisms. Considering the three sources, the general percentage of words chosen over words examined is 1.03%, almost double the previous research, and better yet, more than 21 languages far exceed 1%.

Unless otherwise stated, the analysis that follows refers only to the third source. The second finding is the linguistic families that have large contributions in absolute and relative terms. In Table 3, the contributions per linguistic family are observed, clearly highlighting three families: Bantu, Austronesian, and Romance, among the three they gather more than 50% of all contributions.

The Bantu languages, consisting of Swahili, Zulu, Nyanja, Kinyarwanda, Xhosa, and Shona, are all spoken by millions of people in different parts of Sub-Saharan Africa. All of these languages are used in various domains of everyday life. Swahili is spoken by over 100 million people in countries such as Kenya, Tanzania, Uganda, and parts of the Democratic Republic of Congo. It is considered to have high vitality and is widely used in government, education, media, and business. On the other hand, Austronesian languages are a language family spoken by over 386 million people in many countries across Southeast Asia, the Pacific, and Madagascar. Austronesian languages range from being highly vital to severely endangered. Some Austronesian languages, such as Malay, Tagalog, and Javanese, have high vitality and are widely used in various domains such as education, media, and business.

The representative languages of these families enjoy good health and although none of them is known for being a massive source of neologisms, given their daily use, their speakers are undoubtedly accustomed to quickly creating or adapting neologisms from English, the world's most influential supplier. A great conclusion is that it is worth continuing to mine word lists of these languages in search of more neologisms.

With 9.02%, the third force is the Romance language family, which includes Catalan, Spanish, Portuguese, Romanian, Italian, and French. Incidentally, in the second source, Latin and ancient Greek are in the top ten contributors in absolute terms. It would be interesting if these two languages could become a source of neologisms for the Quechua language, considering that they have been a source of neologisms for practically all European languages. Although Spanish's contribution is not negligible, it is far from being the first contributor, being eighteenth in relative terms (1.22%) and thirteenth in absolute terms (4,554 words). This contrasted fact validates there is no linguistic reason why the Spanish language should be seen as the only source of neologisms for the Quechua language.

Another finding is that the contribution of neologisms from the English language is also small, and this poses a major challenge as it is well known that this language is the most widespread and the one that coins most neologisms; therefore the strategies to introduce neologisms from this language into the Quechua language will have to be reviewed. Surely it will be impossible to avoid the rephonologization of English words. The Chinese language did present a noticeable amount of contributions both in absolute terms (7,667 words) and in relative terms (2.13%); here the usefulness of the improvement of the selection algorithm stands out since with its first version, applied on the first two sources, the Chinese language had no contributions. With the improved algorithm to properly treat diacritics, the contribution of the Chinese language is now notable. Two other widespread languages, Russian and Hindi, had few contributions, and the Arabic language produced even fewer, with only four words, an extremely small number. Surprisingly, Epitran did not support the Korean and Japanese languages, so the third source did not include these languages. However, from the first two sources, neologisms from these languages were obtained, but they were few. It is necessary to review the case of these four languages in depth.

**Table 3.** Proposed neologisms by linguistic family contribution

	Family	Contributions	
		Absolute	Relative
1	Bantu	46,474	24.94%
2	Austronesian	39,422	21.16%
3	Romance	16,799	9.02%
4	Turkic	14,721	7.90%
5	Slavic	9,710	5.21%
6	Indo-Aryan	7,979	4.28%
7	Mandarin	7,667	4.11%
8	Afro-Asiatic	5,562	2.99%
9	Dravidian	5,365	2.88%
10	Hmong-Mien	4,868	2.61%
11	Niger-Congo	4,212	2.26%
12	Indo-European	3,945	2.12%
13	Sino-Tibetan	3,380	1.81%
14	Semitic	3,375	1.81%
15	Germanic	3,285	1.76%
16	Indo-Iranian	2,673	1.43%
17	Austroasiatic	2,301	1.23%
18	Mongolic	1,498	0.80%
19	Tai-Kadai	1,321	0.71%
20	Cushitic	1,065	0.57%
21	Uralic	700	0.38%
	<b>TOTAL</b>	<b>186,322</b>	<b>100.00%</b>

## 9. Future work

There are two major tasks on the horizon: qualitative analysis and diffusion of proposed neologisms among Quechua speakers. This research has focused on quantitative analysis, specifically the number of neologisms generated using a rule-based algorithm. The next crucial step involves assessing the quality of these neologisms through collaboration with native speakers and language experts. Their input is vital to ensure authenticity and cultural relevance. Conducting validation studies and seeking feedback from language communities can refine the selection process and enhance the acceptance of new words. Additionally, conducting long-term impact assessments on the adoption and usage of neologisms can gauge the effectiveness of language revitalization efforts. Monitoring the



integration of new words into daily language use provides insights into the success of the coining process.

Expanding the proposed method to other endangered languages would be valuable in assessing its generalizability and effectiveness in diverse linguistic contexts. Each language may present unique challenges and requirements for coining new words, which warrant further exploration. Although our method holds promise for other languages, further research is necessary to confirm its efficacy. Moreover, to ensure scalability, it is important to normalize the code to facilitate smooth implementation for any language.

As language technology and resources continue to grow, incorporating more comprehensive and diverse data sources could improve the accuracy and coverage of proposed neologisms. Exploring other lexical databases, linguistic corpora, or even crowdsourced platforms could yield additional words for consideration. I would like to improve the method by incorporating additional factors into the rule-based algorithm. For example, we would like to incorporate factors such as the semantic meaning of words and the frequency of use of words.

While the rule-based algorithm used in this research proved effective, machine learning techniques, such as neural networks, could be employed to enhance the selection of neologisms. But again, the lack of data from endangered languages is a barrier to overcome.

Finally, narrowing the scope of the research to specific domains or industries (e.g., technology, environment, arts) could help identify neologisms that address the specific needs and interests of language communities. Tailoring the approach to these specific contexts could lead to more impactful results. However, the lack of free-access authoritative vocabularies and dictionaries from most world languages is another barrier to overcome.

## 10. Conclusions

I discussed the importance of linguistic diversity and how coining new words can promote and preserve endangered languages. As the world becomes increasingly globalized, there is a risk of losing linguistic diversity as people shift to dominant languages. By coining new words and encouraging their usage, we can contribute to revitalizing endangered languages and prevent them from becoming extinct. By preserving linguistic diversity and promoting the usage of endangered languages, we can contribute to preserving cultural heritage and creating a more inclusive and linguistically rich global society. I firmly believe that each language itself should be the primary source for their own neologisms but I also exposed the limitations of that approach.

In this article, I introduce a method for automating the coining of new words for endangered languages. My approach utilizes a rule-based algorithm to search within foreign language lexicons and identify words phonologically similar to those in the target language. I evaluated the method in the Quechua language, finding that it suggests almost 200,000 adoptable words. This figure represents all possible suggestions that meet Quechua's phonotactic constraints; however, I do not expect all of these neologisms to be adopted. The introduction of neologisms always entails a level of artificiality. However, my method does not intend to impose a rigid or disproportionate preference for any particular language, as the algorithm's selection is purely based on phonotactics without cultural bias. Even more, this proposal aims to explore alternative sources beyond the dominant languages to avoid reinforcing historical diglossia.

The selection algorithm is based on phonotactic compatibility, but I agree that cultural relevance and semantic motivation are essential factors. As I proceed, these criteria will be incorporated through engagement with the Quechua-speaking community to ensure the selected neologisms are meaningful and grounded in the Quechua worldview. That

implies qualitative analysis to filter, adapt, and select the most relevant and necessary terms. I will share this proposal not only with decision-makers but also with the broader Quechua-speaking practitioner (most of them IBE teachers) community via the popular website Qichwa2.0<sup>11</sup>, a search engine that indexes the most comprehensive collection of Quechua-Spanish bilingual dictionaries. Our database will be integrated into this platform as a dictionary of neologisms, with a clear disclaimer that this resource is not meant to serve as an authoritative lexicon but as a proposal aiming for wider visibility and consensus. The goal is to support speakers with a rich repository of options, recognizing that only a fraction will ultimately become part of everyday use.

This work has important implications for language revitalization, ensuring new words are phonologically compatible and promoting the creativity of speakers. The method shows promise, but further research is needed to improve accuracy and evaluate effectiveness in other languages. Incorporating more data sources can enhance the accuracy and relevance of suggested neologisms. I hope these findings inform the South American States and encourage support for linguistic bridges among native languages and any other ones.

## References

- Ahlers, Jocelyn & Leanne Hinton. 1999. The issue of “authenticity” in California language restoration. *American Anthropological Association* 30(1), 56–67.
- Ahn, Emily, and Eleanor Chodroff. Voxcommunis: A corpus for cross-linguistic phonetic analysis. Proceedings of the Thirteenth Language Resources and Evaluation Conference, 2022.
- Aikio-Puoskari, Ulla. 2018. Revitalization of Sámi languages in three Nordic countries: Finland, Norway, and Sweden. In Leanne Hinton, Leena Huss, Gerald Roche (eds.), 355–363.
- Akmajian, Adrian, Demers Richard A., Farmer Ann K. & Harnish Robert M. 2010. *Linguistics: An introduction to language and communication*, 6th edn. Cambridge and London: MIT Press.
- Algeo, John, and Adele S. Algeo (eds.). 1991. *Fifty years among the new words: A dictionary of neologisms 1941–1991*. Cambridge: Cambridge University Press.
- Betz, Werner. 1949. *Deutsch und Lateinisch: die Lehnbildungen der althochdeutschen Benediktinerregel*. Bonn, Germany: Bouvier.
- Black, Alan W. 2019. CMU wilderness multilingual speech dataset. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).
- Blair, Heather & Shirley Fredeen. 1995. Do not go gentle into that good night. Rage, rage, against the dying of the light. *Anthropology & Education Quarterly* 26(1), 27–49.
- Camacho, Luis. 2023. A primer on getting neologisms from foreign languages to under-resourced languages. arXiv preprint arXiv:2304.10495
- Cerrón-Palomino, Rodolfo. 1990. Préstamos, elaboración léxica y defensa idiomática. *Allpanchis* 22.36B, 361–392.
- Coronel-Molina, Serafin M. 2007. Language policy and planning, and language ideologies in Peru: The case of Cuzco's High Academy of the Quechua Language (Qheswa simi hamut'ana kuraq suntur). University of Pennsylvania, 2007.
- Coronel-Molina, Serafin M. 2015. Language ideology, policy and planning in Peru. Vol. 161. *Multilingual Matters*. Bristol: Channel View.
- Fellbaum, Christiane. 16 wordnet: An electronic lexical resource. *Oxford handbook of cognitive science*, 301.
- García, Ofelia. 2015. Language policy. In: J. D. Wright (ed.), *International Encyclopedia of the Social & Behavioral Sciences*, 2nd edn., Vol. 13, 353–359.
- Grzega, Joachim. 2003. Borrowing as a word-finding process in cognitive historical onomasiology. *Onomasiology Online* 4.2003, 22–42.
- Haimovich, Gregory & Jan Sześciński. 2018. *A guide to Spanish-Quechua: Language contact phenomena in the Colonial Era*. Faculty of Liberal Arts, University of Warsaw.
- Haspelmath, Martin. 2009. Lexical borrowing: Concepts and issues. In Martin Haspelmath & Uri Tadmor (eds.), *Loanwords in the world's languages: A comparative handbook*, 35–54. Berlin: De Gruyter Mouton.
- Haugen, Einar. 1950. The analysis of linguistic borrowing. *Language* 26(2), 210–231.

<sup>11</sup> <https://www.dic.qichwa.net/#/>

- Huamancayo Curi, Edinson Ysrael. 2017. Neologismos en lenguas originarias: aproximaciones conceptuales y metodológicas. Dirección de Educación Intercultural Bilingüe, Ministerio de Educación de Perú.
- Jackson L. Lee, Lucas F. E. Ashby, M. Elizabeth Garza, Yeonju Lee-Sikka, Sean Miller, Alan Wong, Arya D. McCarthy & Kyle Gorman. 2020. Massively multilingual pronunciation modeling with WikiPron. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, 4223–4228.
- Li, Xinjian, David R. Mortensen, Florian Metze & Alan W. Black. 2021. Multilingual phonetic dataset for low resource speech recognition. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).
- Mortensen, David R., Siddharth Dalmia & Patrick Littell. 2018. Epitran: Precision G2P for many languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Paris, France. European Language Resources Association (ELRA).
- Nahir, Moshe. 1984. Language planning goals: A classification. *Language Problems and Language Planning* 8, 294–327.
- Oksaar, Els. 1996. The history of contact linguistics as a discipline. In Hans Goebel, Herbert Ernst Wiegand et al. (eds.), *Handbücher zur Sprach- und Kommunikationswissenschaft*, Vol 11–12, Berlin, Germany: Walter de Gruyter, 1–12.
- Sager, Juan C. 1990. *A practical course in terminology processing*. Amsterdam/Philadelphia: John Benjamins.
- Sallabank, Julia & Jeanette King. 2021. What do we revitalize? In Justyna Olko & Julia Sallabank (eds.), *Revitalizing endangered languages: A practical guide*, 33–46. Cambridge: Cambridge University Press.
- Sofroniev, Pavel & Çağrı Çöltekin. 2018. Phonetic vector representations for sound sequence alignment. *Proceedings of the Fifteenth Workshop on Computational Research in Phonetics, Phonology, and Morphology*.
- van der Voort, Hein. 2016. Word formation in South American Languages, ed. by Swintha Danielsen, Katja Hannss, and Fernando Zúñiga. *Anthropological Linguistics* 57(3), 340–343.
- Vitevitch, Michael S. & Paul A. Luce. 1999. Probabilistic phonotactics and neighborhood activation in spoken word recognition. *Journal of Memory and Language* 40(3), 374–408.
- Weinrich, Uriel. 1953. Languages in contact. Findings and problems. *Publications of the Linguistic Circle of New York* 1.